Reliability of Telephone and Videoconference Methods of Cognitive Assessment in Older Adults with and without Dementia

Title Page

Title: Reliability of Telephone and Videoconference Methods of Cognitive Assessment in Older Adults With and Without Dementia: a Review

Authors: Matthew B **Hunter**, Natalie **Jenkins**, Clare **Dolan**, Hannah **Pullen**, Craig **Ritchie**, and Graciela **Muniz-Terrera**

Institution (all co-authors): Edinburgh Dementia Prevention, University of Edinburgh, Edinburgh, UK

Corresponding Author: Matthew B Hunter, Matthew.Hunter@ed.ac.uk, Edinburgh Dementia Prevention, University of Edinburgh, BioCube 1, Little France Road, Edinburgh, EH16 4UX. Tel: 0131 651 7828

Keywords: telehealth, teleneuropsychology, telemedicine, MCI, Alzheimer's disease, neuropsychological assessment

Abstract

Background

Telephone and videoconference administration of cognitive tests introduce additional sources of variance compared to in-person testing. Reviews of test-retest reliability have included mixed neurocognitive and psychiatric populations with limited consideration of methodological and statistical contributions.

Objective

We reviewed reliability estimates from comparison studies of older adults with and without dementia, considering test-retest analyses and study methods.

Methods

Medline, Embase, PsycINFO, and Web of Science were systematically searched from 1st January 2000 to 9th of June 2020 for original articles comparing telephone or videoconference administered cognitive instruments to in-person administration in older adults with and without dementia or mild cognitive impairment.

Results

Of 4125 articles, 23 were included: 11 telephone (N=2 dementia cohorts) and 12 videoconference (N=4 dementia cohorts). Telephone administered subtest scores trended in the same direction as in-person with comparable means. Person-level data were scarce. Data on dementia was only available for MMSE, with resulting subtle modality bias. MMSE, SMMSE, Letter Fluency, and HVLT-R in healthy to mild-moderate Alzheimer's disease were particularly reliable for videoconference administration. Other tests show promise but require more observations and comprehensive analyses. Most studies used high-speed stable videoconferencing hardware resulting in a lack of ecological validity for home administration.

Discussion

Remote administration is often consistent with in-person administration but variable and limited at the person/test level. Improved statistical design and inclusion of dementia related cohorts in telephone studies is recommended. Reliability evidence is stronger for videoconferencing but with limited applicability to home administration and severe dementia. Improved reporting of administrative procedures is recommended.

Introduction

Administration of cognitive tests via remote methods (i.e. telephone and videoconference) is likely to accelerate exponentially following the coronavirus disease (COVID-19) pandemic, and extend beyond its mitigation. The pandemic has hampered in person testing and forced researchers to implement remote testing to secure the continuity of existing studies. Whilst it is expected that performance on these tests remains consistent across administration modalities in order to retain reliability, tests should also demonstrate the same reliability in clinical groups who may be differentially affected by remote administration. The feasibility of remote cognitive administration has been previously demonstrated in mixed diagnostic groups [1-5], but a focused review of the data obtained from older adults with dementia is particularly relevant. Thus, a review of the literature examining test-retest reliability estimates derived from comparisons of remote and in-person administered cognitive tests would better inform clinical and research practice in this vulnerable diagnostic group. Additionally, a critical examination of the methodological approaches in the contexts of those reliability estimates is also required in order to increase confidence in, and generalisability of, the findings.

For cognitive tests to be considered valid, they are expected to retain stability across repeated administrations so that differences in scores can be attributed to true change in the individual rather than inconsistency of the measurement. This stability is measured via test-retest analyses. Remote administration may affect test-retest performance by introducing additional sources of variance. For example, unstable audio and sound quality [6], data transmission speed [7], and environmental or psychological variables created by the absence of the researcher [8], have been shown to impact test scores. Furthermore, the generalisability of the estimates may also be affected indirectly through variables that can bias sample selection, such as visual and auditory impairment [9], access to equipment, lack of technical expertise, or apprehension with technology, which can deter engagement and retention [10, 11]. Although numerous cognitive instruments have been designed for telephone administration [12], researchers may need to use modified versions of in-person tests to retain

continuity in study design. Consequently, conclusions drawn from test-retest comparisons may be limited if test modifications are required in order to accommodate remote administration.

The aforementioned sources of variance may become more relevant in certain populations, where age, culture, or diagnosis act as modifiers. Indeed, differential performance across administration modalities between psychiatric and neurocognitive groups is unclear (e.g. [6]). Generalised reliability estimates taken from mixed age and diagnostic populations may be insufficient to apply specifically to old age and dementia. Reliability of cognitive tests across a videoconferencing modality is of particular importance in older adults who are less familiar with internet technologies than younger age groups [13, 14]. Likewise, remote administration is a potential confound to those with dementia, who display difficulties in using and communicating via telephone [15], and who are less likely to use the internet than those without dementia [16]. Thus, an assessment of the current reliability data in dementia populations is advantageous, particularly as remote assessment will be more widely adopted after the cessation of coronavirus restrictions, due its potential advantages in terms of cost saving and reaching underrepresented groups. This is especially relevant to dementia where participation in clinical and non-clinical research remains a major barrier [17].

The feasibility of remote administration of cognitive tests has been widely demonstrated. However, recent reviews have included both neurocognitive and psychiatric disorders [18, 19], or been of videoconference administration only [2-5]. These reviews provide an excellent summary of the feasibility of remote administration in mixed clinical populations but none have focused on dementia. Finally, there has been limited interpretation of the estimates within the methodological contexts of the studies [4] and no review has considered the types of test-retest analyses conducted in the studies. Reliability studies often report estimates derived from bivariate correlations and means testing, both of which do not account for differences at the individual level [20]. Conducting statistical methods at

the paired level (e.g. intraclass correlation), and including an assessment of acceptable levels of variance (e.g. limits of agreement) increases the strength of the estimate [20].

The purpose of the current review is to build upon these previous reviews by focusing on healthy older and dementia-related persons (i.e. Mild Cognitive Impairment (MCI), Alzheimer's disease (AD), and other dementias), taking into account the methodological contexts where the estimates have been derived. To do this, studies which compared standardised cognitive tests during telephone or videoconferencing administration to in-person administration were identified. Reliability estimates are summarised at the test level and a comprehensive qualitative appraisal of the robustness of the reliability estimates is provided by considering variables which directly influence the quality of the estimate (i.e. study design and statistical approaches), the generalisability of the estimate (i.e. sample size, study entry and retention), and performance moderators (i.e. caregiver support). We also report on aspects of validity when reported. The goal is to inform clinicians and researchers on test level reliability in dementia research, limitations of remote administration, and future study design and reporting.

Methods

A systematic search of the published literature was carried out to identify original research articles which have conducted statistical analyses on the reliability and/or validity of standardised cognitive instruments under a remote setting (i.e. telephone or videoconference) in comparison with an in-person setting. The target population was older adults with or without dementia due to Alzheimer's disease. We extended the review to include mild cognitive impairment (MCI) and dementias of any pathology after a pilot search revealed few studies focused on Alzheimer's disease.

The electronic databases Medline (Ovid), Embase (Ovid), PsycINFO (APA), and Web of Science (Clarivate Analytics) were systematically searched on the 9th of June 2020 for articles published from 1st January 2000 to the date of search. The year 2000 was chosen to coincide with rapid development of, and low-cost public access to, videoconferencing technology. Databases were searched using Boolean operators of terms derived from three key concepts, 1) teleservices, 2) neuropsychology/cognition, and 3) assessment/administration. See Table 1 for database specific search terms. The reference lists of articles meeting inclusion criteria were also screened for additional articles.

Database searches were performed by one author (MH). Citations were imported into EndNote X7 and deduplicated using the EndNote parameters; author, year, title, and reference type, ignoring spacing and punctuation. Rayyan QCRI (https://rayyan.qcri.org/welcome) was used to manage the screening process. Four reviewers (MH, NJ, CD, and HP) completed title and abstract screening, and full text review. Two groups of two reviewers independently screened one half of the articles. Conflicts were resolved by consensus agreement between the two reviewers, or by whole group consensus when a decision could not be reached. The same process was applied to reference screening and full text review.

Inclusion criteria were : peer reviewed full-text original research articles using within-group or mixed between-within groups design, English language publications, neurologically healthy adults, MCI,

Alzheimer's disease, or Dementia (of any type), and with age range ≥40 years, in order to capture prodromal or preclinical groups. Exclusion criteria were: non-English language publications or foreign-language adaptations of cognitive assessments, non-original research publications (i.e. abstracts, conference proceedings, letters, reviews, editorials, systematic reviews, meta-analysis, opinion, commentaries, dissertations, and book chapters), studies of remote cognitive assessment without comparison to in-person test scores, intervention studies, case studies or case series, participant age <40 years or no subgroup ≥40 years with quantitative data, studies exclusively using tech-hardware-delivered cognitive applications, studies using only questionnaires or self-report assessment, studies applying between group analyses only, and non-dementia related or medical populations which may confound comparisons (e.g. psychiatric conditions, brain injury, stroke, or eye disorders). As the pilot search found few studies with Alzheimer's disease participants, studies with the latter criterion were included if cases were few and were deemed by the research team not to significantly confound the reliability data.

Data were independently extracted from eligible studies, according to a team piloted proforma, by one of four team members (MH, NJ, CD, and HP) with each extraction quality checked by another team member. Participant data, reliability and validity data, study methods (i.e. administration procedures, cognitive tests and adaptations, testing procedures, and statistical approaches) were extracted. Only results of comparative analyses between telephone/videoconference and in-person administration were obtained, and not that of any wider study analyses.

It should be noted that test-retest reliability reflects the consistency of an instrument to produce the same results across repeated administrations where all other variables, including testing modality, are kept constant. In this review, we refer to test-retest across different modalities, i.e. remote and in-person administrations, as cross-modal test-retest reliability. We comment on within modality test-

retest reliability (i.e. same test repeated twice within the same modality to the same participants) when conducted. Likewise, concurrent validity (i.e. the ability of the test to discriminate between two clinically different populations) is reported where the article assesses the ability of the cognitive instrument to discriminate healthy and non-healthy participants under both modalities.

In this review, Pearson's correlation coefficient r values of <0.4, 0.4-0.6, >0.7, are interpreted as weak, moderate, and strong, respectively [21], and intraclass correlations coefficient (ICC) values of <0.5, 0.5-0.75, 0.75-0.90, and >0.90 are interpreted as poor, moderate, good, and excellent, respectively [22]. Kappa values of <0.60, 0.60-0.80, and >0.80 are interpreted as weak, moderate, and strong agreement, respectively [23]. Other tests of limits of agreement and equivalence are considered acceptable if values fall within original author specified confidence intervals or as judged as acceptable upon inspection of visual plots (e.g. Bland Altman).

Results

Overview

Of 4125 unique articles obtained from database searches, 18 met criteria. A further 11 articles were identified from reference screening, five of which met criteria and were also included. Thus, twenty-three studies were included in this review. Eleven compared in-person administration to telephone administration, and 12 compared in-person administration to videoconferencing administration. The exclusionary process is illustrated in Figure 1.

The twenty-three included studies were comprised of 2166 participants. Eight studies included neurologically healthy adults (n=850), and nine included mixed healthy and dementia-related populations (n=1082). Six studies included analysis of MCI/Alzheimer's disease/dementia participants (n=234); two telephone based [24, 25] and four videoconferencing based [10, 26-28].

Telephone Administration

Overview

Eleven articles compared in-person to telephone administration (Table 2). Ten evaluated cognitive instruments commonly used in-person, whilst one evaluated a telephone-validated battery of cognitive subtests in-person [29]. The studies were primarily composed of neurologically healthy individuals (7 of 11 studies: [29-35] or of a mix of healthy and dementia participants (2 of 11 studies: [36, 37], with mean age typically in the mid-70s to early 80s (10 of 11 studies). Participant cohorts were mostly educated to at least high school level (50-91% of participants) or with a mean of 14 years of education (12.2 - 14.9 years). Only two studies provided analysis of dementia cohorts, one of which was an Alzheimer's disease cohort of mild to severe impairment [25], and the other a validity study of healthy and various dementia participants (67% Alzheimer's disease) [24]. Overall, the proportion of female participants (62.6%) was greater than males, although no study reported sex differences when

analysed. Caucasian ethnicity (81.3%) was marginally higher than the US population where ten of 11 studies originated [38].

Reliability and Validity

Reliability was evaluated in ten of 11 studies, internal consistency in one [37], and concurrent validity in another [24]. The number of subtests and subscales evaluated for cross-modal reliability and validity across studies was large (N=33), leading to a limited number of observations per subtest. As reliability and validity are limited to the scale itself, these are described in turn. However it is important to note that, in general, the statistical coverage of cross-modal test-retest reliability was limited, with only one study applying ICC as a measure of agreement [29], and only one study applying a test of equivalence [34]. Most studies (10 of 11) applied a test of association (Pearson's r) and/or a means difference test to assess reliability. For perspective, of the 33 tests, subtests, or subscales with a reliability estimate, only three were analysed comprehensively with a test of agreement, means difference test, and equivalence test. Table 3 details the coverage of statistical analyses at the subtest/subscale level and illustrates the strength of cross-modal reliability according to the estimates reported in the articles.

Screening Instruments - Three broad screening instruments were examined: the Orientation Memory Concentration test (OMC; [31]), the Telephone Interview for Cognitive Status-modified (TICS-m; [32]), and the Mini Mental State Examination (MMSE). All reported strong associations between in-person and telephone administration. However, none included a test of agreement, resulting in a lack of evidence of reliability at the person level. The MMSE was evaluated in three studies, and all were modified to remove visual and motor items, or were restricted to Orientation questions only [34]. Strong positive associations between the common questions administered across both modalities were reported [25, 37] but means differences and unequal variances indicated inconsistencies of the test [25, 34]. Both healthy and Alzheimer's disease participants exhibited a bias for telephone administration [25, 34], particularly for Orientation and Recall questions. Notably this bias occurred

when the studies conducted the in-person administration in a research facility and the telephone administration at the home of the participant [25, 34]. In contrast, Kennedy et al. [37] conducted both the telephone and in-person administrations of the MMSE at the home of the participant, finding excellent item agreement for the Orientation questions. Thus, there may be a benefit from external cues or familiarity of the home environment during Orientation and Recall questions. Good internal consistency was reported for telephone administered MMSE with a comparable alpha coefficient to in-person administration [37].

Memory – Observations were limited for memory tests/subtests, with a lack of agreement and equivalence testing (Table 3). Data were limited to one observation in healthy participants for the California Verbal Learning Test, CVLT [32], the Hopkins Verbal Learning Test Revised, HVLT-R [30], and the Telephone Interview for Cognitive Status Modified (TICS-m) Word List Learning [29], and two observations for Wechsler Memory Scale (WMS) Logical Memory [29, 34]. The Rivermead Behavioural Memory Test (RBMT) Delayed Story Recall subtest was administered to a mixed population [36]. Thus, no memory test data were available for any dementia related cohorts.

Generally, memory subtests showed mostly moderate associations and comparable means across the 16 subtests/subscales reported, offering positive but limited support for cross-modal reliability. There was limited statistical coverage across tests, with the exception of WMS Logical Memory, signifying a lack of supportive evidence for consistency at the person level. Tests of agreement were limited to the TICS-m and WMS Logical Memory [29], with the TICS-m and WMS Logical Memory Immediate exhibiting weak agreement. However, interpretation may be confounded by a long test-retest delay and absence of modality counterbalancing [29].

A subtle cross-modal difference between learning, immediate and delayed memory was detected. The CVLT exhibited moderate associations and comparable means across all subscales, except for List B, a test of immediate memory for a list of words which had a weak association [32]. Likewise, the TICS-m Word List Learning Immediate subscale was also weakly correlated across modalities and with weak

agreement [29]. No clear modality bias was ascertained. In contrast, list learning and delayed recall tasks were generally better correlated across modalities, regardless of tool (RBMT Story Recall Delayed, WMS logical memory, CVLT List A and Long Recall, and TICS-m word list delayed). Taken together, this suggests greater instability for immediate memory of word lists, a cognitively demanding task, and a greater degree of consistency for repeated exposure (i.e. learning) and delayed memory.

The concurrent validity of the telephone administered Memory Impairment Screen was substantiated in one study, where dementia participants and healthy participants were correctly classified by the scale [24]. Two thirds of the dementia participants had Alzheimer's disease, although the severity of the dementia was unknown. There was some evidence to indicate that the in-person administration had greater sensitivity than the telephone version when specificity was optimized.

Executive Function – Reliability estimates were mostly supportive for Category and Letter Fluency, where moderate to strong correlations with comparable means were reported across modalities in both healthy and mixed populations [29, 30, 32, 34, 36] (Table 3). Weak agreement was noted for Category Fluency where a bias for in-person administration was noted [29]. However, Animal category was used in both modalities, and modality was not counterbalanced, suggesting the bias may have been due to practice effects [29, 39].

The Oral Trail Making Test (OTMT) and Mental Alternation Test (MAT), which are similar verbal tests of number counting and number-letter sequencing, demonstrated comparative means and equivalence in healthy individuals, except in OTMT A where a marginal mean telephone bias was noted [33, 34]. Limits of agreement tests were not conducted for either the OTMT or MAT, meaning consistency at the individual level is unclear.

Digit span reliability findings were variable across studies of healthy individuals [30, 32, 34]. Comparable means and equal variances were recorded in Digit Span Forwards, but weak and moderate associations (r= .36 to .61) were also described, potentially indicating some person level

variation in scores across modalities. Within modality test-retest correlations of 0.60 to 0.83 [40, 41] have been previously reported, potentially signifying greater inconsistency across modalities compared to within. Digit Span Backwards appeared inconsistent across studies, with borderline or dissimilar means and variances recorded [32, 34], despite a significant association [32].

WORLD Backwards showed comparable means between modalities but unequal variances [34], potentially suggesting individual variability across modalities. Relatedly, Kennedy et al [37] noted weak item agreement for letters R and O in the WORLD Backwards task during the MMSE, where it is used as an alternative to the 'serial 7s' task, pointing to a potential source of variability. Alternatively, inconsistency due to restricted range of scoring has been suggested [34], which can amplify small test-retest differences.

No data were available for dementia related participant groups for any of the executive function related test/subtests.

Language – Reliability estimates from the language tests/subtests, the Boston Naming Test (15 Item, BNT-15), Verbal Naming Test, and WAIS Similarities, were limited due to single observations and lack of agreement and variance analyses. All analyses were in healthy participants. Positively, all three test/subtests returned moderate to strong associations between in-person and telephone administration [29, 30, 35]. The BNT-15 also had comparable means [30], providing support of consistency at the group level. However, there was no analysis at the person level. The Verbal Naming test and WAIS Similarities exhibited dissimilar means and agreement, respectively [29, 35], limiting support for stability across modalities.

Visuospatial - One study tested visuospatial abilities in healthy 62-63 year olds via telephone by posting stimuli of the Hands and Object Rotation subtests to the participant [29]. These subtests measure the orientation judgement of line drawings of hands and objects. A problem with administration of the stimuli during the Object Rotation, in the absence of an in-person researcher to organise and issue the stimuli, resulted in partial completion of the test. It is therefore unclear if the

resulting weak association and ICC was due to an effect of telephone administration or logistical issues. Encouragingly for the administration of visual tests across the telephone medium, the Hands subtest showed a strong correlation and strong ICC between telephone and in-person administration.

Time and Order Effects - All studies tested participants across two occasions. The impact of time between testing did not appear to affect the estimates of reliability on inspection of the data, with no discernible patterns in outcome observed between studies whose intervals were under one month, up to three months, or more than three months.

There was some evidence of a modality order effect. Rapp et al [32] tested four groups, two in which participants were administered subtests in the same modality (i.e. in-person & in-person, and telephone & telephone), and two where administrations were cross-modal (i.e. telephone & in-person, and in-person & telephone). The authors found more variable correlations across subtests in the cross-modal groups compared to the same-modality groups, particularly when telephone administration followed in-person administration. Therefore, test stability may decrease when a test is administered across modalities.

Administration Considerations

Factors potentially affecting reliability estimates need to be considered as confounding variables but were inconsistently reported across studies. Test-retest analyses requires tight control of moderating variables, yet over half of studies (N=6) did not describe if administrators were the same or different across modalities. The majority of studies (N=7) did not report pre-test management of external cues in the home environment which could influence responses, such as calendars. Three asked participants to turn off distractions (i.e. TV/Radios/phones) [34-36], whilst only one study explicitly asking participants to remove visual memory cues [36]. The presence of an aid (i.e. partner or caregiver), who could knowingly or unknowingly provide cues, was documented in only two of the 11 studies [25, 36]. The role and influence of the aides were unclear in those studies. Lastly, all studies used the same version of the subtest they administered, or did not indicate otherwise, across both

testing occasions. No study directly assessed alternate form reliability across telephone administration. Thus, reliability estimates are restricted to the specific forms used.

Retention could not be ascertained in four studies [30, 32, 34, 36], but was excellent in the remaining seven studies, signifying the acceptability of telephone based assessments in older adults with and without dementia. To further determine the impact of telephone administration on retention, we examined retention by modality order. Of four studies where telephone assessment followed in-person administration (same day up to three months later), no dropout was observed in three [25, 31, 35], whilst a relatively small dropout of 4% was reported in the fourth [37]. Data from McComb et al [33], who compared four groups within and across modalities, as described earlier, allowed a direct comparison of dropout by modality order. Dropout in groups involving at least one telephone administration (17.2%) was less than the in-person in-person group (30.6%), providing evidence that telephone assessments do not negatively affect retention and thus participant selection bias. Relatedly, it should be noted that the estimates of reliability across all 11 studies may have inherent selection bias. Of the few studies to document exclusion data, Bunker et al. and Lipton et al. reported that 20-25% of participants excluded were because of hearing or language difficulties [24, 30]. Thus, studies may overrepresent healthier individuals. Hearing and language difficulties may impact entry of dementia related participants in particular, although there was insufficient data to consider this here.

Telephone Administration Discussion

Cross-modal test-retest reliability via telephone administration was largely carried out using tests of association and mean differences. Positively, the findings across the large array of subtests examined trended in the same direction as in-person administered tests and produced comparable mean scores. Only a small number of subtests produced weak or inconsistent results. However, given a small number of observations per subtest and the lack of analyses of agreement and equivalence, a definitive conclusion on the strength of cross-modal test-retest reliability remains elusive for almost

all subtests. What is clear is that telephone administered versions of standardised cognitive tests are feasible in older adults and mixed community populations, and provide encouraging findings of stability at the group level that need to be built upon with further analyses at the person level.

The data reviewed here have shown that telephone administration is deliverable to healthy older persons, regardless of gender, and in persons with mild to severe dementia. However, only one study directly assessed reliability in a dementia-related group, that of Alzheimer's disease, investigating one cognitive test (MMSE, [25]). Further evaluation, using a variety of cognitive tests, is required to confirm its reliability in people with dementia. Ethnicity analyses were rare, but did not significantly affect cross-modal reliability when examined [32]. African-Americans were well represented in several studies [24, 32, 37] but other ethnicities (e.g. Hispanics or Asian) were much less represented.

Most subtests were verbal and there was a paucity of visuospatial and motor tasks. This is understandable given logistical and administrative hurdles, but where clinic-based research is a difficulty for participants (e.g. rural dwellers), and in the absence of videoconferencing technology, its utility [42] may outweigh the costs and could be explored further.

Several points arising from the subtest data require further explanation. The MMSE is an extremely common instrument used in old age and dementia research. Based on the data reviewed here, caution is advised when using the MMSE in a mixed home/research facility study design as a beneficial effect of the home environment on subsections of the MMSE may be evident. Furthermore, as studies removed visual and motor questions of the MMSE the findings cannot reliably be extended to the full MMSE. In regard to memory, the general pattern of moderate correlations, comparable means, but weak agreement seen in memory subtests (Table 3) may indicate differences in cross-modal reliability of individual subtests. Alternatively, it may be reflective of the increased levels of individual variability commonly found in memory testing itself [43-46]. It was not possible to determine why there was greater instability for immediate memory of word lists between modalities, but word list encoding is cognitively demanding in older adults [47] and individual variability is a likely explanation, given similar

mean scores. Interestingly, greater stability in scores were observed in repeated telephone-telephone administration compared to repeated in-person or mixed modality groups [32], as well as a non-significant relative telephone bias of telephone administration overall (0.24 (0.11) standard deviations, p=.03). This data may point toward enhanced performance during telephone administration being a possible explanation. Given the limited number of observations, coincidental findings may also be a possibility.

Data from Rapp et al. [32] suggested an effect of modality continuity where an increased risk of variability in scores was observed when administration switches between modalities than when repeated in same domain. This is an important consideration for longitudinal studies or studies applying a mixed approach, and should be considered in study design and analyses. It should be noted that this finding was drawn from between group comparisons, and although groups were matched for age, ethnicity and education, they may have contained unseen biases.

Lastly, retention rates can inform researchers on the impact of telephone assessment compared to in-person assessment on participants' willingness to maintain engagement in repeated or longitudinal assessment, something vital to study design. Retention in the studies was not well documented but from the available evidence telephone assessment did not negatively impact study drop out, including that of dementia-related participants, and may even be beneficial in maintaining participation.

Videoconferencing Administration

Overview

Twelve studies totalling 740 participants compared videoconferencing administration to in-person administration (Table 4). All studies employed both visual and verbal stimuli. Similar to studies of telephone administration, most studies were performed in the USA (N=8). Two were conducted in Australia [7, 48], and two in Canada [49, 50]. Patient groups were reasonably represented with four

studies including AD and/or MCI participant groups (N=154 participants) [10, 26-28]. Mixed healthy and AD/MCI cohorts made up seven of the remaining eight studies [7, 26, 48, 50-53], with one study focused on a healthy cohort [49]. Throughout studies, participants were generally well-educated with a median 14.1 years education. Only five studies reported ethnicity data, two of which recruited heavily from a rural American Indian population (0% Caucasian, [53]; 54% Caucasian, [27]). Of the remaining three studies 91.8% were Caucasian [26, 28, 52]. Other ethnicities had negligible representation.

Reliability and Validity

One study which met inclusion criteria addressed concurrent validity in videoconference administration compared to in-person administration. In that study, an analysis of covariance successfully discriminated MCI/AD and healthy participant groups in the HVLT-R, Letter Fluency, Category fluency, Boston Naming Test (BNT), Digit Span Forwards and Backwards, and Clock Drawing subtests [27]. All other statistical analyses centred on cross-modal reliability.

Statistical coverage was adequately broad across studies, with tests of agreement and/or measures of variance conducted in 23 of the 29 tests/subtests/subscales. Coverage and strength of reliability data is illustrated in Table 5. In general terms, reliability data were largely favourable toward comparable performance between videoconference and in-person administrations. For most subtests there were moderate to excellent associations and agreements, and comparable mean differences and variances. This held for both verbal and visuospatial stimuli. Findings from subtests and any inconsistencies are described below.

Cognitive Screening and Test Batteries – No adverse or contradictory results were found in three screening instruments, the Montreal Cognitive Assessment (MoCA), MMSE, and Standardised MMSE (SMMSE), and in a broad cognitive test battery, the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), providing favourable support for the reliability of these instruments in the videoconference modality. The number of studies per instrument was low (1-2

studies) with the exception of the MMSE which was evaluated in five studies. The MMSE demonstrated excellent levels of agreement across studies, as well as comparable means and acceptable variances and/or limits of agreement. Furthermore, these data were gathered in mixed populations which included 35-77% dementia related participants with mild to severe impairment; [50, 51, 53, 54], as well as a MCI/AD cohort with moderate to normal impairment [26]. Thus, there is strong support for the reliability of the MMSE in the videoconference modality in healthy and dementia-related participants. Similarly, the SMMSE showed good reliability in mixed cohorts (37-60% dementia related participants; [7, 48]) with excellent agreement, comparable means and acceptable limits of agreement. There was no sub-sectional (e.g. Memory, Orientation etc) or item level analysis provided for the MMSE or SMMSE to compare individual sections or questions. The MoCA demonstrated excellent agreement in a sample of mild to severe Alzheimer's disease participants tested several weeks apart [10].

The RBANS, a neuropsychological test battery consisting of Memory, Visuospatial, Attention, Language, and Total Indices, demonstrated moderate (Visuospatial Index) to strong reliability estimates and comparable means across testing modalities in a small mixed sample of healthy and mildly impaired MCI/AD (61%) participants [52].

Executive Function - Tests of executive function showed mostly favourable results with excellent depth of statistical coverage across most subtests. Cross-modal reliability was particularly robust for Letter Fluency as evidenced by consistently strong reliability data across six studies of healthy, mixed, and MCI/AD participants (Table 5) [26-28, 49, 53, 54]. Category Fluency demonstrated favourable reliability with positive findings in tests of agreement, means differences, and variances [26, 27, 53, 54]. The strength of reliability estimates was slightly weaker in comparison to Letter Fluency, with moderate associations and agreement reported, and a small mean difference in favour of in-person administration noted in one study [27]. Some additional evidence of a trend toward in-person bias was seen in two of the studies where non-significant mean differences were observed only after

correction for multiple comparisons [53, 54]. All studies were modality counterbalanced, eliminating practice effects as a cause. Additionally, performance across modalities did not differ based on disease status [27], providing evidence that any bias was unlikely influenced by disease status. Thus, Category Fluency may demonstrate a marginal trend for in-person administration bias.

Reliability data for Digit Span were largely positive, particularly for Digit Span Forwards. A small mean bias for in-person testing was evident in one study [53] but not in others [26, 27, 54]. With moderate to excellent agreement and similar means reported across studies, Digit Span offers acceptable reliability in healthy and dementia-related populations. The Oral Trail Making Test (OTMT) trial A took longer on average in the videoconference administration modality compared to the in-person modality in a study of mixed participants (35% AD/MCI) [53], although agreement was strong, suggesting the videoconference condition may have had a general slowing effect on the sample in this instance. No differences were noted in OTMT B [53].

Memory - Verbal memory as measured by the HVLT-R and Rey Auditory Verbal Learning Test (RAVLT) was, on the whole consistently reliable across subtests and studies. Association and agreement coefficients for the HVLT-R were strong across subscales and across studies, with comparable means and acceptable variances, signifying consistent levels of comparability at the group and person level in mixed [53, 54] and MCI/AD participants [26, 27] with mild to moderate impairment. There was only an inconsistency noted in the HVLT-R Total Recall subscale [54], where an increased mean score was noted in the videoconference modality compared to in-person testing. Interestingly, one other study also noted an increased mean score during the videoconference modality, which was considered non-significant after correction for multiple comparisons ($p<0.04$, [53]). As these studies counterbalanced modalities, the results are unlikely due to practice effects.

Mean scores on the RAVLT were comparable across modalities in a neurologically healthy cohort, with acceptable limits of agreement [49], signifying good reliability at the group level. An absence of a test of agreement analysis prompts the need for further evidence at the individual level.

Language – Cross-modal analyses of language subtests were conducted mostly at the group level, with favourable results from means difference tests in WAIS III Vocabulary [49], BDAE Picture Description, MAE Token Test, and Aural Comprehension of Words and Phrases [28]. The latter subtests were administered to a small group of mild AD participants, demonstrating feasibility of language assessment over videoconference administration. There were no agreement tests with which to assess individual variability across modalities.

The BNT was well examined with excellent statistical coverage across five studies of mixed and MCI/AD participants [26-28, 53, 54]. Findings relate mostly to the BNT short form (i.e. BNT-15) which was used in all but one study [28]. Strong associations/agreement reported across studies indicate good reliability at the individual level. Minor inconsistencies at the group level were noted in the BNT-15, with a significant mean difference across modalities noted in a study of mixed participants (35% MCI/AD, [53]), and unequal variance around the means in a large cohort of mixed participants (41% MCI/AD, [54]). Disease status is an unlikely explanation, as studies of MCI/AD cohorts did not produce similar outcomes [26-28]. Additionally, administration was counterbalanced and testing sessions took place on the same day in these studies, suggesting other unknown factors contributed to these anomalies.

Visuospatial – The MMSE, SMMSE and MoCA, described previously, included visual stimuli or motor components. However, no separate analyses were reported for these items. Several studies administered the Clock Drawing test, WAIS III Matrix Reasoning, and the RBANS Visuospatial Index (i.e. Line Orientation and Figure Copy subtests). Comparable means and acceptable limits of agreement were observed for Matrix Reasoning in healthy participants [49]. The Visuospatial Index of the RBANS revealed moderate agreement and comparable means in a mixed cohort [52]. The Clock Drawing subtest was examined in a number of studies with an unstable pattern of findings; weak to moderate agreement [26, 53, 54], comparable means [27, 53, 54], and both acceptable and unequal variances observed [49, 54]. Weak agreement was recorded in a MCI/AD cohort [26] whilst wide limits

of agreement were reported in a healthy participants [49] signifying that disease status was not a sole contributor toward instability across modalities. The results of the Clock Drawing subtest suggest individual test-retest variability in both healthy and dementia-related participants which may be masked by if analysing the group mean alone.

Administration Considerations

Order effects were well controlled with 11 of 12 studies applying modality order counterbalancing. Modality order was not reported in the twelfth [51]. Inter-rater effects on scoring were controlled for in five studies where rater-modality counterbalancing occurred [49], or where the same rater for each participant was used [10, 26, 50, 52]. There was insufficient data to examine influences of inter-rater and intra-rater effects.

Remote testing conditions were typically in the clinic setting, which is well controlled for environment and technical variables (e.g. noise or data speed). The videoconference modality was performed in a research facility in nine of the 12 studies, where connections between videoconference equipment were typically fast, reliable and stable [7, 26-28, 48, 49, 52-54]. The videoconference condition of the three remaining studies were conducted at the participants' primary residence, using dedicated videoconference equipment [50, 51], or internet enabled videoconference software [10]. The MMSE was the only test that was examined in both settings during the videoconference condition, showing good reliability in both. Videoconference and in-person assessments occurred at the same testing location across studies with the exception of Lindauer et al [10] and McEachren et al [50].

The effects of participant technological illiteracy or lack of expertise, or stimulus administration errors on the part of the participant, were mitigated in most circumstances by the presence of staff or caregiver in the videoconferencing context before and during test (Table 4). In most cases of videoconference administration, whether in a research facility or at the primary residence of the participant, a staff member helped the participant with setting up the equipment and providing writing materials and stimuli before testing. Researchers often held up visual stimuli to the camera

[49, 52]. Stimuli was posted to participants in one home-based study [10]. Staff members or caregivers were typically on-hand during the assessments outside the room [27, 52-54], or inside the room [10, 26, 28, 49-51]. Feedback on how often staff/caregivers aided during administration was mostly unknown but minimal when referenced [53, 54].

Adaptations of cognitive tests were rare, minimal when carried out, and limited to visual stimuli or motor tasks. This included enlargement of visual stimuli [10, 49], replacing tapping for clapping during the Letter A task of the MoCA [10], and additional clarification instructions for the Clock Drawing test [54].

It is unclear how many potential participants were denied entry to studies due to visual or hearing difficulties that would have excluded them from videoconferencing. Although several studies stated vision and hearing difficulties as exclusion criteria, only one study documented a hearing screen for compatibility with videoconference administration, which subsequently excluded one person despite the presence of hearing aids [28]. An accurate indication of retention was unclear in most studies. When it could be determined, four had no evidence of drop out [7, 49, 51, 54]. The status of the others was unclear but presumed to have no dropout [26, 28, 48, 50, 52]. In two studies there was indication of missing data [27, 53] but it was unclear if this was specific to the videoconference or in-person condition, or if this was due to drop out. Lindauer et al [10] reported that 5 of 33 AD participants dropped out of their study. Two were due to technical difficulties, whilst an additional four others did not complete the MoCA during videoconference administration due to frustration or difficulty with comprehension. All participants completed in-clinic MoCA.

Videoconferencing Administration Discussion

The data have shown that videoconference administration is feasible, and more importantly, that acceptable reliability estimates have been evidenced across a variety of subtests. Favourable reliability data were obtained largely from studies of mixed populations, with disease specific analyses in a small number of MCI and AD cohorts. Conversion from MCI to dementia is low in the years following

diagnosis [55], and replication in purer diagnostic cohorts would be beneficial. The data do however evidence good feasibility and reliability of a variety of subtests in mild to severe cognitive impairment due to dementia related conditions.

One major limitation of the cross-modal test-retest reliability estimates is their ecological validity. Most videoconference administrations were conducted in a controlled environment, usually a research facility, where there was dedicated point-to-point videoconferencing hardware with highly reliable and fast connection speeds. An observation also noted in Marra et al [4]. The data provide strong ecological validity for studies employing a satellite clinic, but not when outside of the controlled clinic context, such as the home of the participant. Here there is less control of the environment and technological variables are likely to play a more instrumental role. For instance, Lindauer et al [10] successfully assessed Alzheimer's disease participants using home computer equipment. However, several participants did not complete the assessments due to frustration, a finding not seen in several Parkinson's disease studies [6, 56]. This highlights the need to validate disease-specific cross-modal reliability of cognitive instruments in the home environment.

There was variation in the strength of the cross-modal test-retest reliability findings according to specific tests. The support for cross-modal reliability for the MMSE, SMMSE, Letter Fluency, and HVLT-R was particularly strong, justifying their use in videoconference studies. Item level analysis would be beneficial to confirm consistency across modalities at the subsection level and to inform clinical interpretation. Impairment in MCI and dementia participants in studies examining the MMSE was mild to moderate and it should be noted that participants with severe cognitive impairment due to Alzheimer's disease may have increased difficulty with videoconference administration [57]. Other instruments show promising cross-modal reliability (e.g. MoCA, RBANS, RVLT) but require repeated study with comprehensive reliability analyses to be considered robust to videoconference administration.

The findings from the Category Fluency subtest were modest with subtle evidence of an inclination toward an in-person modality bias. This could not be explained by practice effects due to modality counterbalancing, whilst both the Animal and Fruits and Vegetables versions of the subtest were used, with similar results. Category fluency has previously demonstrated weak test-retest agreement [39], and it is plausible that some individuals may have benefitted from in-person assessment. However, the trend was marginal and Wadsworth et al [27] argue that the actual difference between mean Category Fluency scores, and the resulting small effect size, does not indicate meaningful test variance.

The BNT showed good reliability from three studies of mild to moderate MCI and AD participants, with inconsistencies noted in two mixed population studies. Brearly et al [2] noted a marginal bias for in-person administration, and the reasons for this are unknown. As neither of the studies reviewed here included severe AD participants, cognitive impairment is an unlikely explanation. The BNT-15 has slightly lower test-retest reliability than the full BNT [58] which may be a contributing factor.

The Clock Drawing subtest returned inconsistent results, echoing data from other mixed and clinical groups [2, 4]. Weak agreement was noted in MCI/AD participants [26], with better findings from mixed populations [27, 53, 54]. Whilst this may suggest a differential impact of modality between MCI/AD and healthy participants, Wadsworth et al [27] found no difference of modality when the two participant groups were directly compared. Thus, an effect of cognitive impairment in explaining the difference in performance across studies appears unlikely. Hildebrand et al [49] reported large standard deviations and limits of agreement in healthy participants, indicative of more variable performance of the Clock Drawing test itself. Hildebrand et al noted a modality order effect, evidencing more variable performance in those who had videoconference administration first compared to in-person first. This may have been a contributing factor in other studies, and warrants further examination. Alternatively, test-retest reliability coefficients may vary depending on the scoring system [59], and a restricted range of test scores may have exacerbated small differences.

The videoconference medium involves a level of technical expertise, and the ability to comprehend and follow technical instruction on the participant's behalf. Both healthy and AD participants can require detailed guidance of up to two hours preparation time for setting up equipment and downloading the necessary software when in the home environment [10, 11]. The presence of a caregiver to provide technical help for dementia-related persons should therefore be encouraged, if required. Munro Cullum et al [26, 54] noted that help during the assessment was unlikely to be required even for those with mild to moderate AD. However, these studies were performed in clinic and may not represent the home environment. Lindauer et al [10], who assessed AD participants at their home, noted that participants often had to be encouraged to close curtains, adjust lights, move chairs, and reduce distractions. The authors also advised the use of headphones after some participants reported distress when the caregiver was in discussions with the clinician about personal matters. The role and input of caregiver to external bodies was not always well reported in the reviewed studies. Given that such input can affect performance and mood, caregiver instructions and help received, or lack thereof, during administration should be documented.

The requirements for carrying out cognitive assessments themselves required only minimal adaptation compared to in-person testing, and were unlikely to influence reliability estimates. Issues of stimuli presentation were easily mitigated with stimuli held to the camera for participants to observe. No complications were noted. In one study Lindauer and colleagues posted stimuli to participants – a practice that will incur additional costs in larger scale studies. Scoring was completed during testing in studies, where participants were asked to hold their responses to the camera [26, 27, 53]. Item level analyses or agreement would be beneficial to check for scoring errors between these and traditional methods. Using a tape recorder to verify responses can mitigate the impact of background noise of verbal items [28].

An indication of retention across the videoconference environment is useful for study planning, particularly as it has the potential to reach underrepresented groups. However, retention rates were

rarely stated explicitly. Where it could be ascertained, retention was excellent in studies where videoconferencing was performed in a research facility. Thus, there is good evidence for the use of remote satellite clinics in retaining study participation. Retention in the home environment is less clear, although Lindauer et al [10] did note drop out and test incompletion when videoconferencing was conducted at the home of the participant. It is possible that attending a research facility may increase motivation, and in turn, retention, although self-selection bias from motivated individuals may also be a moderating factor.

General Discussion

This review sought to collate and summarise cross-modal test-retest reliability data from comparative studies of in-person and remote assessment using standardised cognitive tests in healthy older persons and in those with dementia. Increased reliance on remote cognitive assessment following coronavirus 2019 is expected in dementia cohorts, yet studies with these cohorts were few, resulting in limited findings. Nevertheless, the data have shown that remote administration of cognitive tests is feasible in individuals with dementia and provides promising reliability, specifically within the videoconferencing modality. There was little evidence to suggest that videoconference administration differentially affects people with dementia persons compared to neurologically healthy, although there was a paucity of data in those with severe cognitive impairment. Whilst telephone administration remains a viable option for reaching those without videoconferencing equipment, caution should be exercised for telephone administration where data was too limited to conclude on its strengths and weaknesses for dementia participants. We found no studies focused on non-Alzheimer's disease dementias, or preclinical/prodromal cohorts, limiting the findings to mixed cohorts or Alzheimer's disease. The data in the current review, derived largely from mixed healthy and dementia-related populations, supports previous conclusions on the general feasibility of remote cognitive assessment found previously [2, 4], and justifies appropriateness for use in population and community samples.

A strength of the current review is its focus on the statistical methods used in the derivation of reliability estimates. Aldridge et al [20] has recently highlighted the lack of appropriate statistical reporting and methods used in reliability studies, whilst Booth, Murray and Muniz-Terrera [60] have made a call to improve the integrity of data collection and psychometric testing in light of increased remote administration. Our data serve to better inform researchers and clinicians on the strength of the data available, as well as to complement recent scoping and systematic reviews and meta-analyses which provide data focused on telephone and videoconferencing administration across various

diagnostic cohorts [2, 4, 18]. Reliability studies of videoconference administration were much stronger in this regard compared to telephone studies which lacked the statistical analyses at the individual level to robustly evidence cross-modal reliability. Based on the current data, the MMSE, SMMSE, and HVLT-R have offered good reliability evidence during videoconference administration. Other common instruments (e.g. MoCA, verbal fluency) have provided moderate to variable reliability thus far.

In terms of the weak reliability estimates for some individual subtests, and inconsistent findings between studies, detailed in the Results sections, these may be partially due to moderating variables introduced during the remote setting. For instance, the home setting may have been advantageous for Orientation questions of the MMSE. Opposingly, they may be the result of variable test-retest coefficients inherent in the tests themselves. Researchers should consider instrument specific test-retest reliability estimates when interpreting findings within their own studies. The limited number of observations at the test level in the current review may also mask chance occurrences of between study heterogeneity leading to a negativity bias based on minor inconsistencies. Furthermore, detailed reliability data at the individual level is often missing from the psychometric properties of the test [20]. Therefore, the findings from this review should not detract from the use of any or all subtests in future study design, but should encourage consideration of instrument selection, choice of reliability tests, reporting of methods, and for researchers to use these findings to better inform analysis and interpretation of their data.

The key message of this review is awareness of limited reliability data at the test-level in dementia related populations, and of encouragement toward improved study design and data interpretation. In line with Aldridge et al [20] and Booth et al [60] it may be advisable for future reliability studies to include item level analysis, tests of agreement and assessment of variance in order to identify individual variability and items susceptible to moderation during remote settings. Further recommendations include reporting data for clinical subgroups, and investigating remote administration under naturalistic home environments. For telephone studies, exploration of

visuospatial stimuli would enhance collection of cognitive markers for dementia related studies. Future reliability studies are encouraged to promote suitable administrative recording and design. Explicit recording of study exclusion data, retention, and administrative methods would be beneficial, so that an assessment of the confounding factors and generalisability of the results can be considered. Consideration should be given to the mitigation of inter-rater and intra-rater effects. Data on ethnicity and education was not always reported or analysed in the studies reviewed, and should be considered. For all types of studies it is recommended to record methods taken to control environmental variables (e.g. noise, lighting, caregiver role) to account for study biases. It may be of interest to research studies using longitudinal and repeated measures design to highlight the possible order effects identified in the data. That is, more variable data as a result of mixed in-person and remote testing, or the order effect of modality (c.f. [52]). A counterbalanced administrative approach should be implemented if possible. We encourage sensitivity analyses in research studies using mixed administration to investigate the possible impact of modality based fluctuations. It makes intuitive sense that participants who are unfamiliar with cognitive tests may benefit from in-person testing first, where explanation and instruction may be more fluid between researcher and participant compared to telephone or videoconference. This could be explored in future analyses. Whilst we have not focused on telemedicine aspects of clinical care, clinicians should be wary that item level data is scarce and remote environments may have the potential to affect item level responses, and thus qualitative interpretations.

A limitation of the current review was the small number of studies found at the test level with which to draw robust conclusions on the consistency of individual subtests. The heterogeneity of subtests prohibited a meta-analysis and meta-regression which would be particularly informative toward understanding the influence of administration variables on reliability estimates. Additionally, despite using a comprehensive systematic search, we identified a number of articles from reference searches. The number of articles included were not dissimilar to previous systematic searches [2, 4] but suggests the possibility that other articles may have been omitted. Despite these limitations, the strength of

this review has been its consideration of administrative and statistical factors on reliability estimates in studies of dementia, which can inform test selection and study design.

In conclusion, reliability data for remote cognitive assessment is still largely limited but shows promising consistency, particularly for videoconferencing administration, when compared to in-person testing. Further data collection with improved administrative design and comprehensive statistical analyses will serve to resolve nuances and clarify cross-modal reliability further.

Acknowledgments

No funding was acquired for the review.

Conflict of Interest/Disclosure Statement

The authors have no conflict of interest to report.

-

References

[1] Ball C, McLaren P (1997) The tele-assessment of cognitive state: a review. *J Telemed Telecare* **3**, 126-131.

[2] Brearly TW, Shura RD, Martindale SL, Lazowski RA, Luxton DD, Shenal BV, Rowland JA (2017) Neuropsychological Test Administration by Videoconference: A Systematic Review and Meta-Analysis. *Neuropsychol Rev* **27**, 174-186.

[3] Castanho TC, Sousa N, Santos NC (2017) When New Technology is an Answer for Old Problems: The Use of Videoconferencing in Cognitive Aging Assessment. *J Alzheimers Dis Rep* **1**, 15-21.

[4] Marra DE, Hamlet KM, Bauer RM, Bowers D (2020) Validity of teleneuropsychology for older adults in response to COVID-19: A systematic and critical review. *Clin Neuropsychol* **34**, 1411-1452.

[5] Munro Cullum C, Grosch MC (2013) Special Considerations in Conducting Neuropsychology Assessment over Videoteleconferencing. 275-293.

[6] Abdolahi A, Bull MT, Darwin KC, Venkataraman V, Grana MJ, Dorsey ER, Biglan KM (2016) A feasibility study of conducting the Montreal Cognitive Assessment remotely in individuals with movement disorders. *Health Informatics J* **22**, 304-311.

[7] Loh PK, Donaldson M, Flicker L, Maher S, Goldswain P (2007) Development of a telemedicine protocol for the diagnosis of Alzheimer's disease. *J Telemed Telecare* **13**, 90-94.

[8] Luxton DD, Pruitt LD, Osenbach JE (2014) Best practices for remote psychological assessment via telehealth technologies. *Professional Psychology: Research and Practice* **45**, 27-35.

[9] Montani C, Billaud N, Couturier P, Fluchaire I, Lemaire R, Malterre C, Lauvernay N, Piquard JF, Frossard M, Franco A (1996) "Telepsychometry": a remote psychometry consultation in clinical gerontology: preliminary study. *Telemed J* **2**, 145-150.

[10] Lindauer A, Seelye A, Lyons B, Dodge HH, Mattek N, Mincks K, Kaye J, Erten-Lyons D (2017) Dementia Care Comes Home: Patient and Caregiver Assessment via Telemedicine. *Gerontologist* **57**, e85-e93.

[11] Udeh-Momoh CT, de Jager-Loots CA, Price G, Middleton LT (2020) Transition from physical to virtual visit format for a longitudinal brain aging study, in response to the Covid-19 pandemic. Operationalizing adaptive methods and challenges. *Alzheimers Dement (N Y)* **6**, e12055.

[12] Castanho TC, Amorim L, Zihl J, Palha JA, Sousa N, Santos NC (2014) Telephone-based screening tools for mild cognitive impairment and dementia in aging studies: a review of validated instruments. *Front Aging Neurosci* **6**, 16.

[13] Prescott C (2019) Office for National Statistics, pp. 1-11.

[14] Ryan C (2017), ed. Bureau USC American Community

Survey Reports, Washington, DC.

[15] Nygård L, Starkhammar S (2003) Telephone use among noninstitutionalized persons with dementia living alone: mapping out difficulties and response strategies. *Scand J Caring Sci* **17**, 239-249.

[16] Choi NG, Dinitto DM (2013) Internet use among older adults: association with health needs, psychological capital, and social capital. *J Med Internet Res* **15**, e97.

[17] Grill JD, Karlawish J (2010) Addressing the challenges to successful recruitment and retention in Alzheimer's disease clinical trials. *Alzheimers Res Ther* **2**, 34.

[18] Carlew AR, Fatima H, Livingstone JR, Reese C, Lacritz L, Pendergrass C, Bailey KC, Presley C, Mokhtari B, Cullum CM (2020) Cognitive Assessment via Telephone: A Scoping Review of Instruments. *Archives of Clinical Neuropsychology* **35**, 1215-1233.

[19] Geddes MR, O'Connell ME, Fisk JD, Gauthier S, Camicioli R, Ismail Z, COVID-19 ftASoCTFoDCBPf (2020) Remote cognitive and behavioral assessment: Report of the

Alzheimer Society of Canada Task Force on dementia care best practices for COVID-19. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **12**, e12111.
[20] Aldridge VK, Dovey TM, Wade A (2017) Assessing Test-Retest Reliability of Psychological Measures. *European Psychologist* **22**, 207-218.
[21] Dancey CP, Reidy J (2007) *Statistics without Maths for Psychology*, Pearson Education Limited, Edinburgh Gate: Harlow.
[22] Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med* **15**, 155-163.
[23] McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* **22**, 276-282.
[24] Lipton RB, Katz MJ, Kuslansky G, Sliwinski MJ, Stewart WF, Verghese J, Crystal HA, Buschke H (2003) Screening for dementia by telephone using the memory impairment screen. *J Am Geriatr Soc* **51**, 1382-1390.
[25] Newkirk LA, Kim JM, Thompson JM, Tinklenberg JR, Yesavage JA, Taylor JL (2004) Validation of a 26-point telephone version of the Mini-Mental State Examination. *J Geriatr Psychiatry Neurol* **17**, 81-87.
[26] Cullum CM, Weiner MF, Gehrmann HR, Hynan LS (2006) Feasibility of telecognitive assessment in dementia. *Assessment* **13**, 385-390.
[27] Wadsworth HE, Dhima K, Womack KB, Hart J, Jr., Weiner MF, Hynan LS, Cullum CM (2018) Validity of Teleneuropsychological Assessment in Older Patients with Cognitive Disorders. *Arch Clin Neuropsychol* **33**, 1040-1045.
[28] Vestal L, Smith-Olinde L, Hicks G, Hutton T, Hart J, Jr. (2006) Efficacy of language assessment in Alzheimer's disease: comparing in-person examination and telemedicine. *Clin Interv Aging* **1**, 467-471.
[29] Thompson NR, Prince MJ, Macdonald A, Sham PC (2001) Reliability of a telephone-administered cognitive test battery (TACT) between telephone and face-to-face administration. *International Journal of Methods in Psychiatric Research* **10**, 22-28.
[30] Bunker L, Hshieh TT, Wong B, Schmitt EM, Travison T, Yee J, Palihnich K, Metzger E, Fong TG, Inouye SK (2017) The SAGES telephone neuropsychological battery: correlation with in-person measures. *Int J Geriatr Psychiatry* **32**, 991-999.
[31] Dellasega CA, Lacko L, Singer H, Salerno F (2001) Telephone screening of older adults using the Orientation-Memory-Concentration test. *Geriatr Nurs* **22**, 253-257.
[32] Rapp SR, Legault C, Espeland MA, Resnick SM, Hogan PE, Coker LH, Dailey M, Shumaker SA, Group CATS (2012) Validation of a cognitive assessment battery administered over the telephone. *J Am Geriatr Soc* **60**, 1616-1623.
[33] McComb E, Tuokko H, Brewster P, Chou PH, Kolitz K, Crossley M, Simard M (2011) Mental alternation test: administration mode, age, and practice effects. *J Clin Exp Neuropsychol* **33**, 234-241.
[34] Mitsis EM, Jacobs D, Luo X, Andrews H, Andrews K, Sano M (2010) Evaluating cognition in an elderly cohort via telephone assessment. *Int J Geriatr Psychiatry* **25**, 531-539.
[35] Wynn MJ, Sha AZ, Lamb K, Carpenter BD, Yochim BP (2020) Performance on the Verbal Naming Test among healthy, community-dwelling older adults. *Clin Neuropsychol* **34**, 956-968.
[36] Reckess GZ, Brandt J, Luis CA, Zandi P, Martin B, Breitner JC, Group AR (2013) Screening by telephone in the Alzheimer's disease anti-inflammatory prevention trial. *J Alzheimers Dis* **36**, 433-443.
[37] Kennedy RE, Williams CP, Sawyer P, Allman RM, Crowe M (2014) Comparison of in-person and telephone administration of the Mini-Mental State Examination in the University of Alabama at Birmingham Study of Aging. *J Am Geriatr Soc* **62**, 1928-1932.
[38] Roberts AW, Ogunwole SU, Blakeslee L, Rabe MA (2018), ed. Bureau USC, Washington, DC.

[39] Lemay S, Bedard MA, Rouleau I, Tremblay PL (2004) Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *Clin Neuropsychol* **18**, 284-302.
[40] Waters GS, Caplan D (2003) The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, & Computers* **35**, 550-564.
[41] Wechsler D (1981) *Adult Intelligence Scale–Revised*, Psychological Corporation, San Antonio, TX.
[42] Salimi S, Irish M, Foxe D, Hodges JR, Piguet O, Burrell JR (2018) Can visuospatial measures improve the diagnosis of Alzheimer's disease? *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **10**, 66-74.
[43] Ivnik RJ, Smith GE, Lucas JA, Petersen RC, Boeve BF, Kokmen E, Tangalos EG (1999) Testing normal older people three or four times at 1- to 2-year intervals: defining normal variance. *Neuropsychology* **13**, 121-127.
[44] Benedict RHB, Schretlen D, Groninger L, Brandt J (2010) Hopkins Verbal Learning Test – Revised: Normative Data and Analysis of Inter-Form and Test-Retest Reliability. *The Clinical Neuropsychologist* **12**, 43-55.
[45] Paolo AM, Tröster AI, Ryan JJ (1997) Test-retest stability of the California verbal learning test in older persons. *Neuropsychology* **11**, 613-616.
[46] Lo AH, Humphreys M, Byrne GJ, Pachana NA (2012) Test-retest reliability and practice effects of the Wechsler Memory Scale-III. *J Neuropsychol* **6**, 212-231.
[47] Cadar D, Usher M, Davelaar EJ (2018) Age-Related Deficits in Memory Encoding and Retrieval in Word List Free Recall. *Brain Sci* **8**.
[48] Loh PK, Ramesh P, Maher S, Saligari J, Flicker L, Goldswain P (2004) Can patients with dementia be assessed at a distance? The use of Telehealth and standardised assessments. *Intern Med J* **34**, 239-242.
[49] Hildebrand R, Chow H, Williams C, Nelson M, Wass P (2004) Feasibility of neuropsychological testing of older adults via videoconference: implications for assessing the capacity for independent living. *J Telemed Telecare* **10**, 130-134.
[50] McEachern W, Kirk A, Morgan DG, Crossley M, Henry C (2008) Reliability of the MMSE administered in-person and by telehealth. *Can J Neurol Sci* **35**, 643-646.
[51] Grob P, Weintraub D, Sayles D, Raskin A, Ruskin P (2001) Psychiatric assessment of a nursing home population using audiovisual telecommunication. *J Geriatr Psychiatry Neurol* **14**, 63-65.
[52] Galusha-Glasscock JM, Horton DK, Weiner MF, Cullum CM (2016) Video Teleconference Administration of the Repeatable Battery for the Assessment of Neuropsychological Status. *Arch Clin Neuropsychol* **31**, 8-11.
[53] Wadsworth HE, Galusha-Glasscock JM, Womack KB, Quiceno M, Weiner MF, Hynan LS, Shore J, Cullum CM (2016) Remote Neuropsychological Assessment in Rural American Indians with and without Cognitive Impairment. *Arch Clin Neuropsychol* **31**, 420-425.
[54] Munro Cullum C, Hynan LS, Grosch M, Parikh M, Weiner MF (2014) Teleneuropsychology: evidence for video teleconference-based neuropsychological assessment. *J Int Neuropsychol Soc* **20**, 1028-1033.
[55] Bruscoli M, Lovestone S (2004) Is MCI really just early dementia? A systematic review of conversion studies. *Int Psychogeriatr* **16**, 129-140.
[56] Stillerova T, Liddle J, Gustafsson L, Lamont R, Silburn P (2016) Could everyday technology improve access to assessments? A pilot study on the feasibility of screening cognition in people with Parkinson's disease using the Montreal Cognitive Assessment via Internet videoconferencing. *Aust Occup Ther J* **63**, 373-380.
[57] Carotenuto A, Rea R, Traini E, Ricci G, Fasanaro AM, Amenta F (2018) Cognitive Assessment of Patients With Alzheimer's Disease by Telemedicine: Pilot Study. *JMIR Ment Health* **5**, e31.

[58] Calamia M, Markon K, Tranel D (2013) The Robust Reliability of Neuropsychological Measures: Meta-Analyses of Test–Retest Correlations. *The Clinical Neuropsychologist* **27**, 1077-1105.
[59] Shulman KI (2000) Clock-drawing: is it the ideal cognitive screening test? *International Journal of Geriatric Psychiatry* **15**, 548-561.
[60] Booth T, Murray A, Muniz-Terrera G (2020) Are we measuring the same thing? Psychometric and research considerations when adopting new testing modes in the time of COVID-19. *Alzheimer's & Dementia* **n/a**.

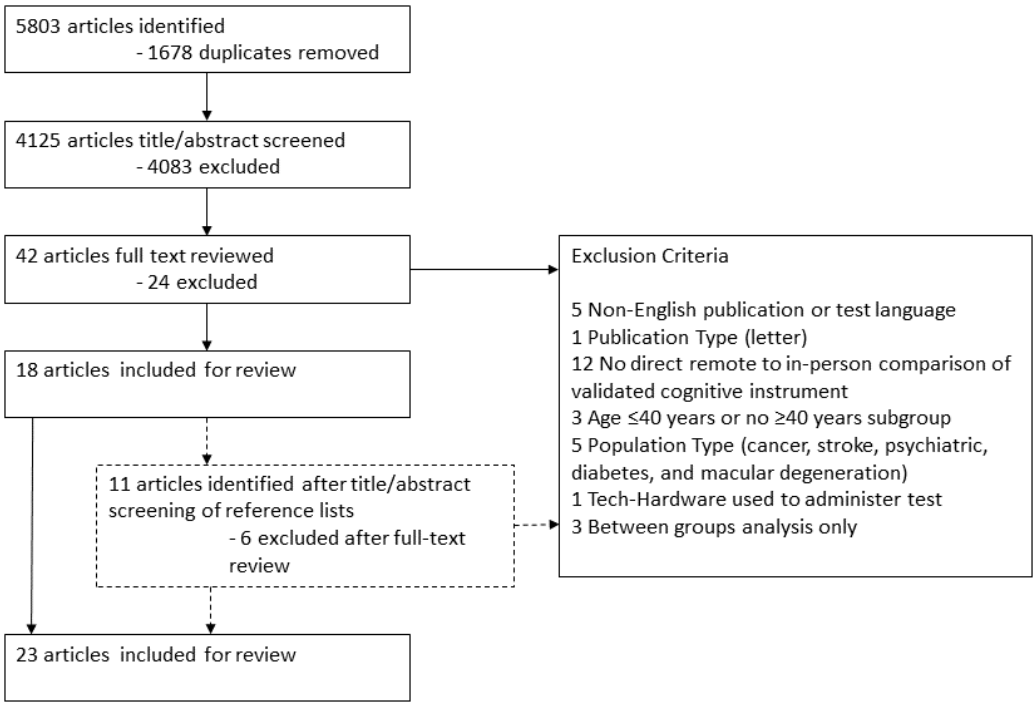

Figure 1 Inclusion Exclusion Flowchart

Table 1 Database specific search terms

Database	Search Terms
Medline, Embase, PsycINFO	(tele*.tw. OR remote.tw. OR video*.tw. OR cyber.tw.) AND (cognition/ OR cognitive ageing/ or neuropsychology/ OR neuropsychological test/) AND (test*.tw. OR assessment.tw. OR administration.tw. OR evaluation.tw.)
Web of Science	(TI=((tele* OR remote OR video* OR cyber) AND (cognition OR "cognitive ageing" OR neuropsychology OR "neuropsychological test") AND (test* OR assessment OR administration OR evaluation))) OR (AB=((tele* OR remote OR video* OR cyber) AND (cognition OR "cognitive ageing" OR neuropsychology OR "neuropsychological test") AND (test* OR assessment OR administration OR evaluation)))

Table 2 Study Details and Cross-Modal Reliability: In-Person versus Remote Telephone Administration

Authors	Participants, N *Mean Age (SD) [Range], years*	Test/Subtest	Modified	MCB	Test-retest Delay	Cross-Modal Reliability Results Favourable	Less Favourable
Dellasega et al [31]	Not Stated, 12 - *81.3 (1.0) [≥65)*	Orientation Memory Concentration Test	No	No	Same Day	*Association Test* - Strong correlation (r=.99)	*None Noted*
Thompson et al [29]	Healthy, 27 - *[62-63]*	Category Fluency Hands Object Rotation NART WAIS similarities Logical Memory: Immediate & Delayed, TICS-m Word List Learning: Immediate & Delayed	Stimuli sent to participants, and novel method of timing used	No	>3 months	*Association Test* - Moderate to strong associations (r=.47 to .95) in 7 of 9 subtests (Animal Fluency, Hands, NART, WAIS Similarities, Logical Memory Immediate & Delayed, Word List Learning Delayed. *Agreement Test (ICC)* - Moderate coefficients (.50 to .72) in 3 of 9 subtests (Hands, NART, and Logical Memory Delayed)	*Association Test* - Weak association in 2 of 9 subtests (Object Rotation and Word List Learning Immediate) *Agreement Test (ICC)* - Poor agreement (.12 to .47) in 6 of 9 subtests (Animal Fluency, Object Rotation, WAIS Similarities, Logical Memory Immediate, Word List Learning Immediate & Delayed) *Means Test* - Marginal to moderate bias for in-person administration compared to telephone in 8 of 9 subtests (0.14 to 0.57 SDs)
Mitsis et al [34]	Healthy, 54 - *79.0 (7.7) [65-97]*	MMSE orientation, WAIS-III Digit Span: Forwards and Backwards, WMS-II Logical Memory: Immediate & Delayed, Letter Fluency, Category Fluency, WORLD backwards, Oral Trail Making Test (OTMT: A & B)*	No	Yes	Same month	*Means Test* - No significant mean differences in 8 of 10 subtests *Equivalence Test* - 7 of 10 subtests were deemed equivalent	*Means Test* - Significant mean difference Digit Span Backwards (telephone bias), and the OTMT A (in-person bias) *Equivalence Test* - Non-equivalent means in MMSE orientation, WORLD backwards**, and Digit Span Backwards
McComb et al [33]	Healthy, 135 - *75.36 (5.85) [65-85]*	Mental Alternation Test	No	Yes	Same month	*Means Test (mixed ANOVA)* - Similar means across modalities, including modality-age interaction (65-75 vs 76-85 years) - Test-retest scores similar for telephone-telephone administrations compared to	*None noted*

						telephone-In-Person administrations	
Rapp et al [32]	Healthy, 105 - *72.4 (5.7) [65-90]*	CVLT (7 subscales), Letter Fluency, Category Fluency, WMS-III Digit Span (Forwards and Backwards), TICS-m	Modified CVLT	Yes	>3 months	*Association Test* - Mostly moderate to strong associations *Means Test* - Similar mean scores in all 12 subtests/subscales*** *Test-Retest (within modality)* - Mean changes in test-retest scores in repeated telephone administration were similar to repeated in-person administration in 10 of 12 subtests/scales - Mean Pearson correlation for repeated telephone-telephone administration (r=0.74 ±0.09) was similar to IP-IP administration (r=0.67 ±0.12)	*Association Test* - Coefficients less stable across modalities compared to within modalities (i.e. telephone-telephone, or In-person-In-person), signifying wavering consistency *Test-Retest (within modality)* - Mean change in test-retest scores in Category Fluency and CVLT Long Delayed Memory were significantly biased toward telephone administration
Bunker et al [30]	Healthy, 50 - 74.9 (4.1)	HVLT-R: Total, Delayed Recall, Discrimination, & Retention, RBANS Digit Span, Category Fluency, Letter Fluency, BNT-15	No	No	Same month	*Association Test* - Modest to strong correlations (r=0.5 to 0.92) in 7 of 8 subtests/subscales *Means Test* - Similar mean scores in all subtests/subscales	*Association Test* - HVLT-R Retention Percentage not significantly correlated between modalities (r=0.27)
Wynn et al [35]	Healthy, 46 - 74.19 (6.5) [61-92]	Verbal Naming Test	No	No	Same week	*Association Test* - Moderate correlation (r=.56)	Means Test - Marginal but significant difference in mean score (1.24 points) between modalities.
Reckess et al [36]	Mixed (dementia & MCI/AD, 27.5%), 225 - [≥70]	RBMT Delayed Story Recall, Category Fluency	No	No	≤3 months	*Means Test* - Similar mean scores in both tests	*None noted*
Kennedy et al [37]	Mixed (dementia 3.3%), 419 - 81.6 (4.8) [≥55]	MMSE	Visuomotor components and orientation item removed (22 point total)	No	≤3 months	*Association Test* - r_s=.69 with the full 30-point MMSE, and r_s=.69 with the sum of the common items to the modified and unmodified MMSEs *Internal consistency*	

						- The telephone (a= 0.845) and in-person (a= 0.763) MMSEs showed good and comparable internal consistency *Item Agreement* - Agreement was strong in 14 of 22 common items (kappa 0.80 to 0.99), and moderate in 3 items (kappa 0.60 to 0.79).	*Item Agreement* - Agreement was weak in 5 of 22 common items (kappa <0.60) (3 items were from delayed recall, and 2 from WORLD backwards)
Lipton et al [24]	Healthy, 273 - *79.1 (5.9)* Dementia (67% AD) - *81.0 (5.7)*	Memory Impairment Screen	No	Yes	≤3 months	*Concurrent Validity* - Administration across both modalities discriminated healthy and dementia participants	*None noted*
Newkirk et al [25]	AD, 53 - *76.5 [56-88]*	MMSE	Visuomotor components and orientation item removed (22 point total)	No	Same month	*Association Test* - *Strong correlation (r=.88)*	*Means Test* - Mean scores significantly higher (1.18 points) in telephone modality - Orientation and recall questions significantly more advantageous during telephone administration, whereas registration questions were significantly more advantageous during in-person administration

MCB – Modality counter-balanced; BNT-15 – Boston Naming Test (15 Item); CVLT – California Verbal Learning Test; HVLT-R - Hopkins Verbal Learning Test-Revised; RBMT - Rivermead Behavioural Memory Test; MMSE – Mini Mental State Examination; TICS-m - Telephone Interview for Cognitive Status - modified

*Selective reminding test administered but not reported here due to trial length differences between modalities (Mitsis et al, 2010)

**Authors suggests non-equivalence is due to "narrow indifference zones" due to restricted range of obtained scores in MMSE orientation and WORLD backwards

***Authors set significance of p value at <0.01. CVLT List A and B, and Digit Span Backward were p<0.05. Mean differences were .20 to .28 SDs

Table 3 Reliability by Cognitive Domain and Cognitive Subtest: Comparison of Telephone and In-Person Administration

Domain	Subtests	Test-Retest Reliability Estimates:				Studies
		Assoc.	Agreement	Means	Equivalence	
Screening Instruments	MMSE	++	NA	+/-	-[1]	Kennedy et al [37]; Mitsis et al [34] (Orientation only); Newkirk et al [25]
	OMC	++	NA	NA	NA	Dellasega et al [31]
	TICS-m (Global Score)	++	NA	+	NA	Rapp et al [32]
Memory	CVLT: List A, Free Recall (short/long), Cued Recall (short/long), Recognition	+	NA	+	NA	Rapp et al [32]
	CVLT: List B	-	NA	+	NA	
	HVLT-R Total Recall, Delayed Recall, Discrimination	++	NA	+	NA	Bunker et al [30]
	HVLT-R Retention Percentage	-	NA	+	NA	
	RBMT Story Recall Delayed	NA	NA	+	NA	Reckess et al [36]
	TICS-m Word List Learning Immediate	-	-	NA	NA	Thompson et al [29]
	TICS-m Word List Learning Delayed	+	-	NA	NA	
	WMS Logical Memory Immediate	+	-	+	+	Thompson et al [29]; Mitsis et al [34]
	WMS Logical Memory Delayed	+	+	+	+	
Executive Functioning	Category Fluency	+/++	-	+	+	Thompson et al [29]; Mitsis et al [34]; Rapp et al [32]; Reckess et al [36]; Bunker et al [30]
	Letter Fluency	++/-[2]	NA	+	+	Mitsis et al [34]; Rapp et al [32]; Bunker et al [30]
	WAIS Digit Span Forwards	+/-	NA	+	+	Mitsis et al [34]; Rapp et al [32]; Bunker et al [30]
	WAIS Digit Span Backwards	+	NA	+/-	-	
	WORLD backwards	NA	NA	+	-	Mitsis et al [34]
	Oral Trail Making Test A	NA	NA	-	+	Mitsis et al [34]
	Oral Trail Making Test B	NA	NA	+	+	
	MAT	NA	NA	+[3]	NA	McComb et al [33]
Language	BNT-15	++	NA	+	NA	Bunker et al [30]
	Verbal Naming Test	+	NA	-	NA	Wynn et al [35]
	WAIS similarities	+	-	NA	NA	Thompson et al [29]
Visuospatial	Hands	++	+	NA	NA	Thompson et al [29]
	Object Rotation	-	-	NA	NA	Thompson et al [29]
Premorbid IQ	NART	++	+	NA	NA	Thompson et al [29]

Association (bivariate correlation): - weak, + modest, ++ strong
Agreement (intraclass correlation): - poor, + moderate, ++ strong, +++ excellent
Means (means difference test): - significant difference, + non-significant difference
Equivalence test: - outwith limits, + within limits
+/- indicates mixed results found across studies

NA: not applicable or no test conducted
1 equivalence test conducted only in MMSE Orientation
2 Associations weak to strong (r=0.33 to 0.95) reported
3 No main effect of modality via a mixed ANOVA

Table 4 Study Details and Cross-Modal Reliability: In-Person versus Remote Videoconference Administration

Authors	Participants, N *Mean Age (SD) [Range], years*	Test/Subtest	MCB	Aide Present	Test-retest Delay	Cross-Modal Reliability Results Favourable	Less Favourable
Hildebrand et al [49]	Healthy, 29 - *68 (8) [≥60]*	RAVLT: Immediate Recall, Short Delay Recall, Long Delay Recall, and Learning, Brief Test of Attention, Matrix Reasoning, Vocabulary, COWAT, Clock Drawing^	Yes	Yes (inside room)	Same month	*Means Test* - Similar mean scores in 8 of 9 subtests/subscales *Limits of Agreement* - Vocabulary and RAVLT Short Delay recall, had the narrowest limits of agreement (-3.07 to +3.13, and -3.65 to +5.31 respectively)	*Means Test and Limits of Agreement* - Clock Drawing exhibited a notable mean difference (-1.93, SD 10.07), with wide limits of agreement (-22.07 to +18.21) - When videoconferencing modality was administered first, wider limits of agreement were more common compared to when in-person modality was administered first
Grob et al [51]	Mixed, <27[1] - *72.5 (2.8)*	MMSE	unknown	Yes (inside room)	Same week	*Agreement Test (ICC)* - Excellent ICC value of .95	*None noted*
Loh et al [48]	Mixed (AD 37.5%), 16 - *82 [72-95]*	SMMSE	Yes	Not stated	Unknown	*Means Test and Limits of Agreement* - Similar mean score (0.20, SD 1.50) and narrow 95% limits of agreement (–3.20 to 2.80)	*None noted*
Loh et al [7]	Mixed (AD/dementia 60%), 20 - *79 [67-89]*	SMMSE	Yes	Not stated	Same week	*Agreement Test (ICC)* - ICC was good (0.89, 95% CI 0.75 - 0.96) *Means Test and Limits of Agreement* - Similar mean scores, and acceptable limits of agreement	*None stated*
McEachern et al [50]	Mixed (MCI/AD/dementia 77%), 71 - *72 (11) [42-89]*	MMSE	Yes	Yes (inside room)	≤3 months	*Means test and Limits of Agreement* - Similar mean scores, and acceptable levels of agreement	*None noted*
Munro Cullum et al [54]	Mixed (MCI/AD 41%), 202 - *68.5 (9.5) [46-90]*	MMSE, HVLT-R, Letter Fluency, Category Fluency, BNT-15, Digit span: Forwards and Backwards, Clock Drawing^^	Yes	Yes (outside room)	Same day	*Agreement Test (ICC)* - Moderate to excellent agreement across tests/subtests (ICCs 0.55 to 0.91) *Means, Variances, and Limits of Agreement* - Similar means and variances (Bradley-Blackwood procedure), and/or acceptable limits of agreement (Bland-Altman plots) in 6 of 8 subtests	*Means, Variances, and Limits of Agreement* - Mean HVLT-R score was significantly higher in the videoconference modality, and there were unequal variances in the BNT-15 despite comparable means

Galusha-Glasscock et al [52]	Mixed (MCI/AD 61%), 18 - *69.7 (7.8) [58-84]*	RBANS: Total, Immediate Memory, Visuospatial, Language, Attention, Delayed Memory	Yes	Yes (outside room)	Same day	*Agreement Test (ICC)* - ICCs for all 6 indices were moderate to good (0.59 to 0.90). Mean ICC=0.80 *Means Test* - Mean scores for all indices were similar	*None noted*
Wadsworth et al [53]	Mixed (MCI/AD 35%), 84 - *64.9 (9.7) [46-88]*	MMSE, HVLT-R: Total, Delayed Recall, & Retention, Letter Fluency, Category Fluency, BNT-15, Digit Span: Forwards and Backwards, Clock Drawing, Oral Trail Making Test: A and B	Yes	Yes (outside room)	Same day	*Agreement Test (ICC)* - ICCs across all 12 subtests/subscales were good to excellent (0.65 to 0.93) *Means Test* - Mean scores were similar on 9 of 12 subtests/subscales	*Means Test* - Small significant mean biases for in-person modality were recorded in Digit Span Forward, Oral Trail Making Test A, and BNT-15
Munro Cullum et al [26]	MCI/AD, 33 - *73.3 (6.9) [51-84]*	MMSE, HVLT-R: Total Recall, Delayed Recall, Retention, and Recognition, RBANS Digit Span, Category fluency, Letter Fluency, BNT-15, Clock Drawing	Yes	Yes (inside room)	Same day	*Association and Agreement Tests* - All subtests/subscales (excluding Clock Drawing) had moderate to strong associations (r=0.55 to 0.89), and moderate to good agreement (ICC=0.54 to 0.88) *Means and Variances Test (Bradley-Blackwood procedure)* - Similar means and variances reported across all subtests/subscales	*Agreement Test* - Agreement (75.8%) was weak (kappa=0.48) in the clock drawing test
Vestal et al [28]	AD, 10 - *73.9 (3.7) [68-78]*	BDAE Picture description, BNT, MAE Token test, Letter Fluency	Yes	Yes (inside room)	Unknown	*Means Test* - Similar mean ranks across all subtests	*None noted*
Lindauer et al [10]	AD, 33 - *71.6 (11.6) [51–96]*	MoCA^”	Yes	Yes (inside room)	Same month	*Agreement Test (ICC)* - Agreement was excellent (ICC=0.93) - Visuospatial/executive subsection had excellent agreement (ICC=0.86), and Clapping Task had moderate agreement (kappa=0.69)	*None noted*
Wadsworth et al [27]	Healthy, 119 - *66.1 (9.2)*	HVLT-R: Total, and Delayed Recall, Letter Fluency,	Yes	Yes (outside room)	Same day	*Means Test (ANCOVA)* - Similar mean scores in healthy and MCI/AD groups in 7 of 8	*Means Test (ANCOVA)*

	MCI/AD, 78 - *72.7 (8.4)*	Category fluency, BNT-15, Digit span Forwards and Backwards, Clock Drawing				subtests/subscales. Small effect sizes (Cohen's d .007 to .202) indicate only a small amount of variance attributable to testing modality *Concurrent Validity (ANCOVA)* - Tests discriminated healthy and MCI/AD groups across modalities	- Small but significant mean bias for in-person administration in Category Fluency

BDAE - Boston Diagnostic Aphasia Examination (auditory response version); BNT (15) - Boston Naming Test (15 Item); COWAT – Controlled Oral Word Association Test; HVLT-R – Hopkins Verbal Learning Test – Revised; MAE - Multilingual Aphasia Examination; MoCA – Montreal Cognitive Assessment; RBANS – Repeatable Battery for Assessment of Neuropsychological Status; SMMSE – Standardised Mini Mental State Exam

1 Study had two groups totalling 27 participants. Results here include only the videoconference-in-person group of unknown number. Percentage of dementia participants unknown.

^visual stimuli larger than normal

^^words changed between administrations for memory components

"participants clapped instead of tapped during Letter A task

Table 5 Reliability by Cognitive Domain and Cognitive Subtest: Comparison of Videoconference and In-Person Administration

Domain	Subtests	Test-Retest Reliability Estimates:				Studies
		Assoc.	Agreement	Means	Variance	
Broad Tests	RBANS					Galusha-Glasscock et al [52]
	- Visuospatial	NA	+	+	NA	
	- Immediate Memory, Language, Attention, Delayed Memory, Total	NA	++	+	NA	
	Montreal Cognitive Assessment	NA	+++	NA	NA	Lindauer et al [10]
	SMMSE	NA	++	+	+	Loh et al [48]; Loh et al [7]
	MMSE	++	+++	+	+	Grob et al [51]; McEachern et al [50]; Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al [53]
Executive	Letter Fluency	++	++/+++	+	+	Hildebrand et al (2004); Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al [53]; Wadsworth et al [27]; Vestal et al [28]
	Category Fluency	+	+	+/-	+	Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al [53]; Wadsworth et al [27]
	Digit Span Forward	++	+/++	+/-	+	Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al [53]; Wadsworth et al [27]
	Digit Span Backward	NA	+	+	+	Munro Cullum et al [54]; Wadsworth et al [53]; Wadsworth et al [27]
	Oral Trail Making Test					Wadsworth et al [53]
	- Test A	NA	++	-	NA	
	- Test B	NA	++	+	NA	
Memory	Hopkins Verbal Learning Test Revised					Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al (2016); Wadsworth et al [27]
	- Total Recall	++	++	+/-	+	
	- Delayed Recall	++	++	+	+	
	- Retention Percentage	++	++	+	+	
	Rey Auditory Verbal Learning Test					Hildebrand et al [49]
	- Immediate recall, short delay recall, long delay recall, & learning	NA	NA	+	+	
Language	Boston Naming Test	++	++	+/-	+/-	Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al [53]; Wadsworth et al [27]; Vestal et al [28]
	Vocabulary Subtest (WAIS III/WASI)	NA	NA	+	+	Hildebrand et al [49]
	Picture Description (BDAE)	NA	NA	+	NA	Vestal et al [28]
	Token Test (Multilingual Aphasia Examination)	NA	NA	+	NA	Vestal et al [28]
	Aural Comprehension of Words and Phrases	NA	NA	+	NA	Vestal et al [28]
Visuo-spatial	Clock Drawing Test	NA	+/-[1]	+	+/-	Hildebrand et al [49]; Munro Cullum et al [26]; Munro Cullum et al [54]; Wadsworth et al [53]; Wadsworth et al [27]
	Matrix Reasoning (WAIS III/WASI)	NA	NA	+	+	Hildebrand et al [49]

Association (bivariate correlation): - weak, + modest, ++ strong
Agreement (ICC, intraclass correlation): - poor, + moderate, ++ strong, +++ excellent
Means (means difference test): - significant difference, + non-significant difference
Variance (tests of variances; limits of agreement, Pitman test, or Bradley-Blackwood procedure): - outwith limits, + within limits

+/- indicates mixed results found across studies
NA: not applicable, no test or association or mean difference conducted
1 Cohen's Kappa was weak (0.48; Munro Cullum, 2006)